

CFM 2529 03



日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2002年 1月31日

出 願 番 号

Application Number:

特願2002-023953

[ ST.10/C ]:

[ JP2002-023953 ]

出 願 人

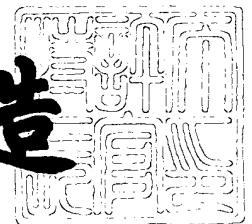
Applicant(s):

キヤノン株式会社

2002年 3月22日

特 許 庁 長 官  
Commissioner,  
Japan Patent Office

及 川 耕 造



出証番号 出証特2002-3019194

【書類名】 特許願

【整理番号】 4597004

【提出日】 平成14年 1月31日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 3/00

【発明の名称】 プローブ設計方法及び情報処理装置

【請求項の数】 50

【発明者】

【住所又は居所】 東京都大田区下丸子3丁目30番2号 キヤノン株式会社  
社内

【氏名】 吉井 裕人

【特許出願人】

【識別番号】 000001007

【氏名又は名称】 キヤノン株式会社

【代理人】

【識別番号】 100076428

【弁理士】

【氏名又は名称】 大塚 康徳

【電話番号】 03-5276-3241

【選任した代理人】

【識別番号】 100112508

【弁理士】

【氏名又は名称】 高柳 司郎

【電話番号】 03-5276-3241

【選任した代理人】

【識別番号】 100115071

【弁理士】

【氏名又は名称】 大塚 康弘

【電話番号】 03-5276-3241

【選任した代理人】

【識別番号】 100116894

【弁理士】

【氏名又は名称】 木村 秀二

【電話番号】 03-5276-3241

【先の出願に基づく優先権主張】

【出願番号】 特願2001- 55465

【出願日】 平成13年 2月28日

【手数料の表示】

【予納台帳番号】 003458

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 0102485

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 プローブ設計方法及び情報処理装置

【特許請求の範囲】

【請求項 1】 未知の核酸断片とハイブリダイゼーション反応させて遺伝子解析を行うためのプローブに用いられる塩基配列を設計するプローブ設計方法であって、

ターゲットの塩基配列に基づいて得られる複数の部分塩基配列を各ノード上に配置するツリーを生成する生成工程と、

所望のノードに関して、これと接続する前記ツリー上の経路に存在するノードが示す部分塩基配列に基づいて、該所望のノードが表す部分塩基配列のプローブとしての適性を評価する評価工程と、

前記評価工程による評価結果に基づいてプローブとして使用すべき部分塩基配列を決定する決定工程と

を備えることを特徴とするプローブ設計方法。

【請求項 2】 前記生成工程における複数の部分塩基配列は、前記ターゲットの塩基配列の相補的な塩基配列から得られる部分塩基配列であることを特徴とする請求項 1 に記載のプローブ設計方法。

【請求項 3】 前記生成工程における複数の部分塩基配列は、前記ターゲットの塩基配列から得られる部分塩基配列であり、

前記決定工程は、前記評価工程の評価結果に基づいて部分塩基配列を選択し、選択された部分塩基配列の相補的な塩基配列をプローブとして使用すべき部分塩基配列として決定することを特徴とする請求項 1 に記載のプローブ設計方法。

【請求項 4】 前記生成工程は、前記ターゲットの塩基配列に基づいて得られる全ての部分塩基配列を分類するツリーを生成することを特徴とする請求項 1 乃至 3 のいずれかに記載のプローブ設計方法。

【請求項 5】 前記評価工程は、前記所望のノードが表す部分塩基配列の中央付近に、ターゲットに対する特異性が変化する塩基配列が存在する場合に、プローブとして適切であるとする評価関数を導入することを特徴とする請求項 1 乃至 3 のいずれかに記載のプローブ設計方法。

【請求項 6】 前記評価工程は、前記経路に存在する各ノードについて、ノードに対応する部分塩基配列の前記ターゲットにおける出現数に基づいてエントロピーを算出し、該算出されたエントロピーの減少状態に基づいて評価を行なうことを特徴とする請求項 1 乃至 3 のいずれかに記載のプローブ設計方法。

【請求項 7】 前記評価工程は、前記所望のノードに対応する部分塩基配列の中央付近における前記エントロピーの変化量を重視する評価関数を導入することを特徴とする請求項 6 に記載のプローブ設計方法。

【請求項 8】 前記決定工程は、前記評価工程において前記評価関数によって求めた値が所定値を超えるノードに対応する部分塩基配列をプローブとして決定することを特徴とする請求項 5 または 7 に記載のプローブ設計方法。

【請求項 9】 前記決定工程は、前記エントロピーの変化量が所定値を超えるノードの部分塩基配列をプローブとして決定することを特徴とする請求項 6 に記載のプローブ設計方法。

【請求項 10】 前記複数の部分塩基配列を前記ターゲットの塩基配列に対する特異性によってグループ分けするグループ化工程を更に備え、

前記決定工程は、前記評価工程の評価結果に基づいて各グループよりプローブとして用いる部分塩基配列を決定することを特徴とする請求項 1 乃至 9 のいずれかに記載のプローブ設計方法。

【請求項 11】 前記複数の部分塩基配列を前記ターゲットの塩基配列に対する特異性によってグループ分けするグループ化工程と、

前記グループ化工程で得られたグループのうち、プローブとして適した特異性を有するグループを選択する選択工程とを更に備え、

前記決定工程は前記選択工程で選択されたグループのそれぞれより、前記評価工程による評価結果に基づいてプローブとして用いる部分塩基配列を決定することを特徴とする請求項 1 乃至 9 のいずれかに記載のプローブ設計方法。

【請求項 12】 前記選択工程は、情報的に完全に独立で必要十分なグループのみを選択することを特徴とする請求項 11 に記載のプローブ設計方法。

【請求項 13】 前記選択工程は、少なくとも解析対象の全てのターゲットに対して特異性のないグループを排除することを特徴とする請求項 11 に記載の

プローブ設計方法。

【請求項 1 4】 前記評価工程は前記選択工程で選択されたグループ内の部分塩基配列について評価を行なうことを特徴とする請求項 1 1 に記載のプローブ設計方法。

【請求項 1 5】 前記ターゲットは複数の塩基配列パターンを含み、  
前記グループ化工程は、前記複数の塩基配列パターンのそれぞれとの反応の有無が同じ部分塩基配列を同一グループに所属させることを特徴とする請求項 1 0 乃至 1 4 のいずれかに記載のプローブ設計方法。

【請求項 1 6】 前記ツリーにおいて、ノードの接続によって表わされる部分塩基配列の塩基配列順序が、前記ターゲット中の塩基配列順序と一致していることを特徴とする請求項 1 乃至 1 5 のいずれかに記載のプローブ設計方法。

【請求項 1 7】 前記ツリーにおいて、ノードの接続によって表わされる部分塩基配列の塩基配列順序が、前記ターゲット中の対応する部分塩基配列の塩基配列順序をその中央部が先頭となるように並べ替えられていることを特徴とする請求項 1 乃至 1 5 のいずれかに記載のプローブ設計方法。

【請求項 1 8】 前記評価工程は、予め指定された範囲の長さの部分塩基配列のみを評価することを特徴とする請求項 1 乃至 1 7 のいずれかに記載のプローブ設計方法。

【請求項 1 9】 前記評価工程は、予め指定された範囲の融解温度条件に適合する部分塩基配列のみを評価すること特徴とする請求項 1 乃至 1 7 のいずれかに記載のプローブ設計方法。

【請求項 2 0】 前記決定工程は、予め指定された範囲の長さの部分塩基配列の中から、前記評価工程による評価結果に基づいてプローブとしての部分塩基配列を決定することを特徴とする請求項 1 乃至 1 7 のいずれかに記載のプローブ設計方法。

【請求項 2 1】 前記決定工程は、予め指定された範囲の融解温度条件に適合する部分塩基配列の中から、前記評価工程による評価結果に基づいてプローブとしての部分塩基配列を決定することを特徴とする請求項 1 乃至 1 7 のいずれかに記載のプローブ設計方法。

【請求項 2 2】 未知の核酸断片とハイブリダイゼーション反応させて遺伝子解析を行うためのプローブに用いられる塩基配列を設計するプローブ設計方法であって、

ターゲットとなる塩基配列に基づいて得られる特定の長さの部分塩基配列を分類する部分塩基配列ハッシュ表を生成する生成工程と、

前記塩基配列ハッシュ表に存在する部分塩基配列に対して、その塩基配列に基づいてプローブとしての適性を評価する評価工程と、

前記評価工程による評価結果に基づいて、プローブとして使用する部分塩基配列を決定する決定工程と

を備えることを特徴とするプローブ設計方法。

【請求項 2 3】 前記生成工程における部分塩基配列は、前記ターゲットの塩基配列の相補的な塩基配列から得られる部分塩基配列であることを特徴とする請求項 2 2 に記載のプローブ設計方法。

【請求項 2 4】 前記生成工程における部分塩基配列は、前記ターゲットの塩基配列から得られる部分塩基配列であり、

前記決定工程は、前記評価工程の評価結果に基づいて部分塩基配列を選択し、選択された部分塩基配列の相補的な塩基配列をプローブとして使用すべき部分塩基配列として決定することを特徴とする請求項 2 2 に記載のプローブ設計方法。

【請求項 2 5】 前記生成工程は、異なる長さの部分塩基配列に対応して複数のハッシュ表を生成することを特徴とする請求項 2 2 乃至 2 4 のいずれかに記載のプローブ設計方法。

【請求項 2 6】 前記評価工程は、部分塩基配列の中央付近に、ターゲットに対する特異性が変化する塩基配列が存在する場合に、プローブとして適切であるとする評価関数を導入することを特徴とする請求項 2 2 乃至 2 4 のいずれかに記載のプローブ設計方法。

【請求項 2 7】 前記ターゲットが複数存在し、

前記評価工程は、前記複数のターゲットの複数の塩基配列間で相互に塩基配列が異なる特異位置を求め、前記ハッシュ表に登録された部分塩基配列中における該特異位置の位置に基づいてプローブとしての適性を評価することを特徴とする

請求項 2 2 乃至 2 5 のいずれかに記載のプローブ設計方法。

【請求項 2 8】 前記評価工程は、プローブとしての適性を評価するために、前記特異位置が塩基配列中の中央部に位置するか否かを調べることを特徴とする請求項 2 7 に記載のプローブ設計方法。

【請求項 2 9】 前記決定工程は、前記評価工程で求めた評価関数の値が所定値を越えたプローブを選択することを特徴とする請求項 2 6 に記載のプローブ設計方法。

【請求項 3 0】 前記複数の部分塩基配列を前記ターゲットに対する特異性によってグループ分けするグループ化工程を更に備え、

前記決定工程は、前記評価工程の評価結果に基づいて各グループよりプローブとして用いる部分塩基配列を決定することを特徴とする請求項 2 2 乃至 2 9 のいずれかに記載のプローブ設計方法。

【請求項 3 1】 前記複数の部分塩基配列をターゲットに対する特異性によってグループ分けするグループ化工程と、

前記グループ化工程で得られたグループのうち、プローブとして適した特異性を有するグループを選択する選択工程とを更に備え、

前記決定工程は前記選択工程で選択されたグループのそれぞれより、前記評価工程による評価結果に基づいてプローブとして用いる部分塩基配列を決定することを特徴とする請求項 2 2 乃至 2 9 のいずれかに記載のプローブ設計方法。

【請求項 3 2】 前記選択工程は、情報的に完全に独立で必要十分なグループのみを選択することを特徴とする請求項 3 1 に記載のプローブ設計方法。

【請求項 3 3】 前記選択工程は、少なくとも解析対象の複数のターゲットに対して特異性のないグループを排除することを特徴とする請求項 3 1 に記載のプローブ設計方法。

【請求項 3 4】 前記評価工程は前記選択工程で選択されたグループ内の部分塩基配列について評価を行なうことを特徴とする請求項 3 1 に記載のプローブ設計方法。

【請求項 3 5】 前記ターゲットは複数の塩基配列パターンを含み、  
前記グループ化工程は、前記複数の塩基配列パターンのそれぞれとの反応の有



無が同じ部分塩基配列を同一グループに所属させることを特徴とする請求項 3 0 乃至 3 4 のいずれかに記載のプローブ設計方法。

【請求項 3 6】 前記評価工程は、予め指定された範囲の融解温度条件に適合する部分塩基配列のみを評価すること特徴とする請求項 2 2 乃至 3 5 のいずれかに記載のプローブ設計方法。

【請求項 3 7】 前記決定工程は、予め指定された範囲の融解温度条件に適合する部分塩基配列の中から、前記評価工程による評価結果に基づいてプローブとしての部分塩基配列を決定することを特徴とする請求項 2 2 乃至 3 5 のいずれかに記載のプローブ設計方法。

【請求項 3 8】 未知の核酸断片とハイブリダイゼーション反応させて遺伝子解析を行うためのプローブに用いられる塩基配列を設計するプローブ設計方法であって、

ターゲットとなる塩基配列データより得られる複数の部分塩基配列のリストを分類する判別ツリーを生成する生成工程と、

前記判別ツリーに存在するプローブ候補に対して、プローブとしての良し悪しを評価する評価工程と、

前記評価工程の評価結果に基づいて使用すべきプローブを選択する選択工程とを備えることを特徴とするプローブ設計方法。

【請求項 3 9】 前記ターゲットとなる塩基配列データは、検体に存在する可能性のある全ての塩基配列を含むことを特徴とする請求項 3 8 に記載のプローブ設計方法。

【請求項 4 0】 前記ターゲットとなる塩基配列データは、検体に存在する可能性のある全ての塩基配列と、特定の塩基配列とを含むことを特徴とする請求項 3 8 に記載のプローブ設計方法。

【請求項 4 1】 前記評価工程は、前記特定の塩基配列の部分配列で構成されるプローブ候補について評価を行なうことを特徴とする請求項 4 0 に記載のプローブ設計方法。

【請求項 4 2】 未知の核酸断片とハイブリダイゼーション反応させて遺伝子解析を行うためのプローブに用いられる塩基配列を設計するプローブ設計方法

であって、

ターゲットとなる塩基配列データより特定の長さの複数の部分塩基配列のリストを分類する部分塩基配列ハッシュ表を生成する生成工程と、

前記部分塩基配列ハッシュ表に存在するプローブの候補に対して、プローブとしての良し悪しを評価する評価工程と、

前記評価工程の評価結果に基づいて使用すべきプローブを選択する選択工程とを備えることを特徴とするプローブ設計方法。

【請求項 4 3】 前記ターゲットとなる塩基配列データは、検体に存在する可能性のある全ての塩基配列を含むことを特徴とする請求項 4 2 に記載のプローブ設計方法。

【請求項 4 4】 前記ターゲットとなる塩基配列データは、検体に存在する可能性のある全ての塩基配列と、特定の塩基配列とを含むことを特徴とする請求項 4 2 に記載のプローブ設計方法。

【請求項 4 5】 前記評価工程は、前記特定の塩基配列の部分配列で構成されるプローブ候補について評価を行なうことを特徴とする請求項 4 4 に記載のプローブ設計方法。

【請求項 4 6】 請求項 1 乃至 4 5 のいずれかに記載のプローブ設計方法を実現する情報処理装置。

【請求項 4 7】 請求項 1 乃至 4 5 のいずれかに記載のプローブ設計方法をコンピュータに実現させるためのプログラム。

【請求項 4 8】 請求項 1 乃至 4 5 のいずれかに記載のプローブ設計方法をコンピュータに実現させるためのプログラムを格納する記憶媒体。

【請求項 4 9】 請求項 1 乃至 4 5 のいずれかに記載のプローブ設計方法を用いて決定された塩基プローブを含む DNA マイクロアレイ。

【請求項 5 0】 請求項 1 乃至 4 5 のいずれかに記載のプローブ設計方法で決定された塩基プローブを用いた遺伝子検査装置。

【発明の詳細な説明】

【0 0 0 1】

【発明の属する技術分野】

本発明は複数の核酸プローブを配置したマイクロアレイの設計支援に好適な情報処理装置及びプローブ設計方法に関する。

#### 【0002】

##### 【従来の技術】

従来より、特開平10-272000号公報または特開平11-187900号公報にあるように、DNAマイクロアレイを用いて遺伝子の発現やシーケンスを決定するシステムが提案されている。このシステムは、あらかじめ試料とハイブリダイズさせる塩基配列断片（プローブ）を設計する必要がある。良いプローブセットを設計することができれば、非常に高い確率で試料の中に存在する塩基配列の断片に対する大量の情報が得られることが特徴である。

#### 【0003】

##### 【発明が解決しようとする課題】

しかしながら、解析対象の試料（ターゲット）の中に存在する可能性のある塩基配列断片は、非常に多様であり、実験系が異なると全く異なる。この場合、全く異なるプローブを設計する必要がある。従来は、このプローブ設計を経験に基づいて人手で行なっていたが、大量の塩基配列が決定されつつある今では、人手でいちいちプローブ設計を行うことは、事実上不可能となりつつある。

#### 【0004】

##### 【課題を解決するための手段】

本発明は、上記の課題に鑑みてなされたものであり、解析対象となるターゲットの塩基配列に応じて、解析に適切なプローブを自動的に選択することを可能とし、プローブ設計を効果的に支援することを目的とする。

#### 【0005】

上記の目的を達成するための本発明によるプローブ設計方法は、  
未知の核酸断片とハイブリダイゼーション反応させて遺伝子解析を行うための  
プローブに用いられる塩基配列を設計するプローブ設計方法であって、  
ターゲットの塩基配列に基づいて得られる複数の部分塩基配列を各ノード上に  
配置するツリーを生成する生成工程と、  
所望のノードに関して、これと接続する前記ツリー上の経路に存在するノード

が示す部分塩基配列に基づいて、該所望のノードが表す部分塩基配列のプロープとしての適性を評価する評価工程と、

前記評価工程による評価結果に基づいてプロープとして使用すべき部分塩基配列を決定する決定工程とを備える。

#### 【0006】

また、上記の目的を達成するための本発明の他の態様によるプロープ設計方法は、

未知の核酸断片とハイブリダイゼーション反応させて遺伝子解析を行うためのプロープに用いられる塩基配列を設計するプロープ設計方法であって、

ターゲットとなる塩基配列に基づいて得られる特定の長さの部分塩基配列を分類する部分塩基配列ハッシュ表を生成する生成工程と、

前記塩基配列ハッシュ表に存在する部分塩基配列に対して、その塩基配列に基づいてプロープとしての適性を評価する評価工程と、

前記評価工程による評価結果に基づいて、プロープとして使用する部分塩基配列を決定する決定工程とを備える。

#### 【0007】

##### 【発明の実施の形態】

以下、添付の図面に基づいて本発明の好適な実施形態を説明する。

#### 【0008】

##### <第1実施形態>

##### 〔装置及び動作の概要〕

以下の実施形態では、いわゆるDNAマイクロアレイを用いた核酸配列解析システムに用いる最適なオリゴヌクレオチドプロープを計算機を用いて設計するプロープ設計方法とそれを実行する装置を説明する。

#### 【0009】

図2は、本実施形態によるプロープ設計方法が適用される情報処理装置の構成を示すブロック図である。本実施形態では、外部記憶装置201、中央処理装置(CPU)202、メモリ203、入出力装置204から構成される情報処理装置に、以下に説明するプロープ設計方法を実行させる。このような構成の情報処

理装置としては、例えばパーソナルコンピュータが挙げられる。

【 0 0 1 0 】

外部記憶装置 2 0 1 は、本実施形態によるプローブ設計方法を実現するプログラムや、ターゲットの塩基配列データ及びパラメータを保持する。また、本実施形態によって導かれたプローブ配列を保持するのにも使用される。中央処理装置（CPU） 2 0 2 は、プローブ設計のプログラムを実行したり、すべての装置の制御を行ったりする。

【 0 0 1 1 】

メモリ 2 0 3 は中央処理装置（CPU） 2 0 2 が実行するプログラム、及びサブルーチンやデータを一時的に記録する。外部記憶装置 2 0 1 に格納されたプローブ設計方法の制御プログラムは、メモリ 2 0 3 にロードされ、中央処理装置 2 0 2 により実行される。入出力装置 2 0 4 は、ユーザとのインタラクションを行う。多くの場合、以下に説明するプローブ設計方法を実現するプログラム実行のトリガはこの入出力装置を介してユーザが出すことになる。また、ユーザが結果を見たり、プログラムのパラメータ制御をこの入出力装置を介して行う。

【 0 0 1 2 】

図 1 は、第 1 の実施形態によるプローブ設計方法の処理手順を説明するフローチャートである。本実施形態のプローブ設計方法処理概要を説明すると、まず、ステップ S 1 0 1 において、外部記憶装置 2 0 1 からターゲットの塩基配列データを取得する。次にステップ S 1 0 2 において、ステップ S 1 0 1 で取得した塩基配列データに基づいて判別ツリーを作成する。判別ツリー及びその作成については後述する。そして、ステップ S 1 0 3 において、ステップ S 1 0 2 で作成した判別ツリーを外部記憶装置 2 0 1 に格納する。

【 0 0 1 3 】

ステップ S 1 0 4 では、ステップ S 1 0 3 で格納した判別ツリーを用いて、ツリー上に表現されている各プローブ（部分塩基配列）の評価を行なう。すなわち、判別ツリーの中に存在するプローブ候補の中からプローブとしてふさわしいものを選択するための評価を行なう。ステップ S 1 0 5 では、上記評価の結果をもとに、最終結果としての最適プローブセットを選択する。ステップ S 1 0 6 では

最適なプローブを出力する。出力先としては、外部記憶装置 2 0 1、入出力装置 2 0 4 に含まれる表示装置或いはプリンタ等が挙げられる。

#### 【 0 0 1 4 】

以下、各処理について詳細に説明する。

#### 【 0 0 1 5 】

[ 判別ツリーの生成について (ステップ S 1 0 1 ~ S 1 0 3) ]

図 3 は、ステップ S 1 0 1 で取得する塩基配列データの例を示す図である。図 3 にある塩基配列情報は H L A または M H C と呼ばれる遺伝子群の一部である。この部分は、人間の抗体生成に関連する遺伝子をコーディングしており個体差がある。図 3 に類似しているが完全に同一ではない塩基配列があるのは、このことを意味している。この H L A に関して現在人類で確認されている遺伝子タイプは 1 0 0 種類以上あり、この種類が異なると生体間移植の際に致命的な拒絶反応がおこる可能性が高い。現在では、患者と臓器提供者の D N A をシーケンスし、完全に塩基配列を決定することによって、この H L A のタイピング (分類) を行っている。この作業を適切な D N A マイクロアレイを用いて行うことによって、簡単かつ高速に行うことができるようになる。

#### 【 0 0 1 6 】

図 4 は D N A マイクロアレイ上のハイブリダイゼーションの様子を示した図である。生体内ではほとんどの場合、D N A 塩基配列は 2 重らせん構造をしていて、その 2 本鎖の間の結合は塩基間の水素結合で実現されている。一方、R N A 塩基配列は 1 本で存在する場合が多い。塩基の種類は D N A の場合は A C G T の 4 種類、R N A の場合は A C G U の 4 種類であり、それぞれ水素結合ができる塩基対は A - T (U)、G - C のペアとなっている。以下では、説明のため、D N A 塩基配列に関するプローブの作成を例に挙げる。

#### 【 0 0 1 7 】

ハイブリダイゼーションとは、1 本鎖状態の塩基配列分子同士がある部分の相補的な塩基配列を介して部分的に結合して 2 本鎖状態を形成することをいう。本実施形態で想定している反応は、図 4 の上側の基板にくっついた塩基配列 (プローブ配列) の方が下側の試料中にある塩基配列分子より短い。よって、試料中に

存在する塩基配列分子がプローブ配列を含む場合は、このハイブリダイゼーション反応はうまくいき、試料中のターゲット塩基配列分子はトラップされることとなる。

#### 【 0 0 1 8 】

但し、プローブ配列の全ての領域が相補的な場合のみ、ハイブリダイゼーションが可能なのではなく、一部にペアを形成しない部分があっても、ターゲット塩基配列分子はトラップされることがある。特に、プローブの末端だけの塩基が結合できないような場合は、ハイブリダイゼーションを起こしてしまう可能性が高く、ターゲット塩基配列の正確な選別が出来ない。従って、末端だけが異なるプローブ配列セットは、ハイブリダイゼーションを使った実験には不適格といえる。すなわち、塩基配列中の中央部にターゲットの塩基配列と特異となる配列が存在するものが好ましい。また、ハイブリダイゼーション反応の強度はプローブの塩基配列の長さが長いほど高くなる。よって、理想的にはDNAマイクロアレイ上に配置されるプローブ配列はハイブリダイゼーション強度が似ているもの、すなわち含まれる塩基数が近い塩基配列を選ぶほうが良い。

#### 【 0 0 1 9 】

上述のように、最終出力であるプローブ塩基配列は、試料中の部分配列と相補的な塩基配列である必要がある。そこで、本実施形態では、ステップ S 1 0 2 の判別ツリー作成処理において、予め全ての部分塩基配列を用意する。この様子を図にしたのが図 5 である。

#### 【 0 0 2 0 】

図 5 はターゲットとなる塩基配列から得られる部分塩基配列（プローブ）を示す図である。図 5 で列挙している塩基配列群は、図 3 の DRB1\*0101 というタイプの配列の部分配列群の一部である。図 5 のように、末端を 1 塩基ずつシフトすることによって、全く異なる部分塩基配列が得られ、この全てがプローブとハイブリダイゼーション反応を行う配列候補となる。あるタイプの対象配列数が  $n$  ベースだったとすると、その部分塩基配列は全部で  $n - 1$  個存在することとなる。本実施形態では全ての部分塩基配列を分類する判別ツリーを生成するが、予め所定数の深さを設定し、所定数以上の塩基を有する部分塩基配列について判別ツリー

を生成するようにしてもよい。

#### 【 0 0 2 1 】

次に、この全ての部分配列を判別／分類する判別ツリーを作成する。図 6 A は判別ツリーを説明する図である。ツリーを構成する全てのノードは、4 つの子ノード (A, T, G, C) を持つ。それぞれ 4 つの子ノードは、注目している位置の塩基配列が A T G C のいずれかで分類される。まず、一番簡単な例として、注目している位置を 1 つずつずらす方法を説明する。

#### 【 0 0 2 2 】

図 6 B は判別ツリーにおける各ノードへの部分塩基配列の分類例を示す図である。ルートノードで、各部分塩基配列の左端の塩基配列を見る。そして、その塩基がどの種類かで部分塩基配列を分類する。例えば、図 5 に示した H L A の DRB1 \*0101 由来の塩基配列ならば、図 5 の 2 番の部分塩基配列は A のノード 6 1 1 に、1, 3, 5 番の部分塩基配列は G のノード 6 1 2 に、4 番の部分塩基配列は C のノード 6 1 3 にそれぞれ分類される。そして、1, 3, 5 番の部分配列は、更に深いノードで細分されることとなる。具体的には、1 番目の部分塩基配列の次のノードは A であり、3 番目の部分塩基配列の次のノードは C であり、5 番目の部分塩基配列の次のノードは G である。よって、それぞれ、ノード 6 1 4、ノード 6 1 5、ノード 6 1 6 に格納されることになる。

#### 【 0 0 2 3 】

以上の様な処理により、全てのターゲットの全ての部分塩基配列について各ノードへの登録を行なって、判別ツリーを作成する。この判別ツリー作成処理では、図 5 に示したような部分配列に展開したデータを全て対象として判別ツリーを作成するので、作成された判別ツリーの上のどこかに任意の部分塩基配列が登録されることとなる。上記例のように位置を左から一つずつずらす方法で判別ツリーを作成した場合には、ある部分塩基配列の登録されたノードをルートノードへ向けて逆に辿ると、登録された塩基配列の左側の部分がわかることとなる。

#### 【 0 0 2 4 】

図 7 及び図 8 は判別ツリー上のあるノードに登録された部分塩基配列の様子を示す図である。例として、図 3 に示した 4 種類の 5 0 ベースからなる塩基配列が



試料に入っていると仮定する。図 6 A で示した判別ツリーの中で、C A G の順でノードを辿っていくと、図 6 A のノード 6 0 1 には、図 7 の表の右側に示した情報が格納されている。すなわち、ノード 6 0 1 には、DRB1\*0101…「NULL」、DRB1\*04011…「21」、DRB1\*07011…「10,37」、DRB1\*15011…「21,37」が格納されている。この情報によれば、DRB1\*0101にはC A Gの部分塩基配列は存在せず(NULL)、DRB1\*04011には2 1番目の場所にC A Gが存在し、DRB1\*07011には1 0番目と3 7番目、DRB1\*15011には2 1番目にそれぞれC A Gが存在することが示されている。

#### 【 0 0 2 5 】

図 7 で示したノードに接続される子ノードの中でTとなるノード、すなわち図 6 A のノード 6 0 2 の内容を示したのが図 8 である。このノード 6 0 2 に対応する部分塩基配列はC A G Tであり、これにマッチングする部分配列の位置は、DRB1\*07011の1 0番目の部分配列のみである。

#### 【 0 0 2 6 】

以上示してきた例は、塩基の種類を左から1つずつずらした位置で見てきた例であり、ノードの順序と塩基配列とが一致している。しかしながら、この位置はアルゴリズムによって変わってもよい。ハイブリダイゼーション反応の場合、中心の塩基配列の相違が重要である。よって、判別ツリーを作成する際に、注目する塩基配列の位置を中心から周辺へ移していくように並べ替えることが一つの有力な方法としてあげられる。

#### 【 0 0 2 7 】

例えば、図 7 の左からC→A→Gとう並びの部分塩基配列に相当するようなノードを、「真中のA」→「右のG」→「左のC」というふうに見ていくことがその方法にあたる。この場合、左から1つずつ見るのではなく、真中、右端、左端の順番で見ていくことになる。従って、左から見る方法だと図 7 のノード 6 0 1 は{ルートノード→Cのノード→Aのノード→Gのノード}という位置に存在するが、中心から周辺へ見る方法だと、{ルートノード→Aのノード→Gのノード→Cのノード}という位置に存在することになる。このことは、注目する塩基配列(上記例ではC A G)の位置を変えると図 6 A においてノード 6 0 1 がノード

603 ( $A \rightarrow G \rightarrow C$ ) に変わることの意味する。この場合、ノード601だけが  
変わるのではなく、判別ツリー全体の構造が変化する。

#### 【0028】

以上のような並べ替えを一般化して示せば、 $n$  個の塩基配列について、 $n$  が奇  
数の場合は  $(n+1)/2$  を、 $n$  が偶数の場合は  $n/2$  を  $a$  とした場合に、「 $a, a+1, a-1,$   
 $a+2, a-2, \dots$ 」というように並べ替えて判別ツリーを構成することになる。も  
ちろん、中心、左端、右端の順番で見えていっても問題ない。

#### 【0029】

以上説明してきた通り、判別ツリー作成処理で判別ツリーが作成されるが、こ  
こで注意するのは、ノードの数は指数関数的に増えていくので、ノードの節約を  
行わないと必要となるコンピュータリソースが不足することである。具体的には  
、想定する試料（ターゲット）中にノードに対応する塩基配列が存在しなくなれ  
ば、その子ノードの伸張はストップさせるという処理が必要となる。この処理に  
よって、ある以上の深さを超えると、増加するノードは、定常的になる。例えば  
、判別すべき塩基配列のタイプの数を  $T$ 、それぞれの全部分配列を  $N$  とすると、  
最悪のケースでも  $T \times N$  個のノードのみが増えることとなる。

#### 【0030】

##### 〔プローブの評価方法について（ステップ S104）〕

次に図1のプローブ評価処理（S104）を説明する。前述した判別ツリーの  
全てのノードは試料中に存在しうる部分塩基配列のどれかをコーディングしてい  
るので、判別ツリーの全てのノードはプローブの候補となりうる。しかし、ハイ  
ブリダイゼーション反応の条件よりプローブとして有効に働く部分塩基配列は限  
られてくる。このプローブとしての評価をするのが以下に説明するプローブ評価  
処理である。

#### 【0031】

まず、チップ上に載るプローブ群はハイブリダイゼーション強度が似ているほ  
うがよい。よって、あらかじめプローブの塩基配列の数を決定する方法が考えら  
れる。この場合、判別ツリーのある深さのノードのみがプローブ候補となる。も  
ちろん、この塩基配列の数に幅を持たせてプローブ候補を絞ってもよい。いうま

でもなく、この場合はある深さからある深さの範囲のノードがプローブ候補となる。

#### 【 0 0 3 2 】

また、厳密に言えば、塩基配列の数が同じでもハイブリダイゼーション強度は異なる。これを計算する方法として、例えば、特開平 8 - 3 1 7 7 9 0 「オリゴリボヌクレオチド／オリゴデオキシリボヌクレオチド相補塩基対構造の安定性解析方法」が提案されている。この方法によれば、塩基配列が決まればそのハイブリダイゼーション強度は推定できる。従って、各ノードに対応する塩基配列から各ノードのハイブリダイゼーション強度を推定し、この推定値を元に、判別ツリー上のノード群に絞込を行なう方法もある。但し、この場合、例えば文献 “Predicting DNA duplex stability from the base sequence” (Proc. Natl. Acad. Sci. USA Vol. 83, pp. 3746-3750, June 1986 Biochemistry) に記載の方法で、判別ツリー上のノードの塩基配列より、塩基配列の融解温度を求めることになる。

#### 【 0 0 3 3 】

本実施形態の非常に優れた点の 1 つは、判別ツリーが任意の長さの部分塩基配列をコーディングするので、ある範囲の長さやある範囲の融解温度を持ったプローブ候補を選ぶ処理を非常に高速に行なえる点にある。

#### 【 0 0 3 4 】

さて、上述の絞込により、例えばある特定の長さ（例えば 2 0 ベース）の部分配列を全て含むプローブを設計しても良いのであるが、全てのターゲットに含まれる部分塩基配列には特異性が全くなく、プローブとして意味がない。すなわち、このプローブを使って各ターゲットとハイブリダイゼーション反応を行なっても、何の情報も得ることはできない。よって、判別ツリーのノードによって特定される部分塩基配列の特異性を測る必要がある。この特異性の尺度として通常用いられるのがエントロピーである。

#### 【 0 0 3 5 】

例えばターゲットとなる塩基配列タイプが全部で  $t$  タイプあり、あるノードに含まれるタイプ毎の部分塩基配列の数が  $N_1, N_2, N_3, \dots, N_t$  であるとする、そのノードのエントロピーは次の式で計算されることとなる。

【0036】

【数1】

$$entropy = \frac{1}{\sum_{i=1}^t N_i} \cdot \sum_{i=1}^t \left[ N_i \left\{ \log \left( \sum_{i=1}^t N_i \right) - \log(N_i) \right\} \right]$$

【0037】

もちろん、ここで説明したエントロピー以外の方法で塩基配列の特異性を測ってもよい。

【0038】

さて、上記の式によって算出されたエントロピーを用いて最適なプローブを探すことになるが、その方法を以下に説明する。図9は、プローブとして典型的に最適なエントロピーの推移をグラフに示した図である。グラフの横軸はノードの深さを示し、この例では15ベースの長さの塩基配列を調べた結果となっている。判別ツリーのノードを辿っていくと、通常、エントロピーは減少していく。例えば、図6Aのノード601よりノード602の方が特異性が高く、エントロピーは減少する。図6Aのノード601、602が図7、図8のような部分塩基配列のセットを保持するとすると、ノード602はDRB1\*07011に固有の部分配列に相当し、エントロピーは0となる。

【0039】

図9のグラフで8番目の塩基と10番目の塩基の場所でエントロピーの減少が見られる。つまり、このノードのコーディングする部分塩基配列の8番目と10番目があるタイプセットに特異的な塩基配列であることがグラフからわかる。図9にあるように、中央付近で急激なエントロピー減少があるノードがプローブとして最適となる。

【0040】

本実施形態の非常に優れた点の1つは、このエントロピー減少の様子がノードを深さ方向に辿ることによって、自然と把握することができ、この解析によって、確実に中心位置での塩基特異性を持ったプローブを選べることにある。

## 【0041】

以上の様なプローブの選択処理を実現するための具体的なプローブ評価関数としては、中心位置にエントロピー減少があると高いスコアになるようなものであればよい。例えば次に示す評価関数を使用することが可能である。

## 【0042】

## 【数2】

$$\text{評価スコア} = (\text{エントロピーの減少量}) * (1 / (\text{中心位置からの距離} + 1))^n$$

但し、この方法では、予めプローブの長さを想定しなければならない。なぜならば、中心位置はプローブの全長を決定して初めて決定されるからである。

## 【0043】

予めプローブの長さを制限せずに評価を行なう方法として、判別ツリー作成ステップでも述べた中心から周辺へ塩基のバリエーションを見ていく方法が挙げられる。この方法で判別ツリーを作成した場合、判別ツリーの最初の方で、急激なエントロピー減少があった方がよく、それに対応した評価関数を用意することとなる。

## 【0044】

例えば、次の評価関数を使用することが可能である。

## 【0045】

## 【数3】

$$\text{評価スコア} = (\text{エントロピーの減少量}) * (1 / (\text{ルートノードからの距離}))^n$$

この方法だと任意の長さのプローブ設計に対応できることとなる。

## 【0046】

以上のような評価関数を導入してプローブの評価を行ない、その評価値を各ノードに格納しておく。

## 【0047】

図10は本実施形態によるプローブの評価処理（ステップS104）とプローブセットの選択処理（ステップS105）の詳細を示すフローチャートである。ステップS201では、上述したように、塩基数（ノード深さ）や塩基配列に基づく融解温度により、判別ツリー上よりプローブ候補の絞込を行なう。そして、

ステップ S 2 0 2 において、上記絞込によって得られたプローブ候補のツリー上のパスにある各ノードについて、〈数 1〉によりエントロピーを計算し、各ノードに格納する。そして、ステップ S 2 0 3 において、プローブ候補のノードについて、〈数 2〉或いは〈数 3〉に示した評価関数を用いて評価値を算出し、これを当該ノードに格納する。

## 【 0 0 4 8 】

なお、図 1 0 において、上記評価関数を用いたプローブの評価は、所定長の或いは所定長範囲の部分塩基配列を抽出して得られたプローブ候補、或いは所定範囲の融解温度を持ったプローブ候補について行うようにしているがこれに限らない。評価を判別ツリー上の全てのプローブについて行い、以下に説明するプローブセットの選択処理時に所定長範囲或いは所定融解温度範囲で絞込を行うようにしてもよい。

## 【 0 0 4 9 】

〔プローブセットの選択処理について（ステップ S 1 0 5）〕

次に図 1 のプローブセット選択処理（ステップ S 1 0 5）を説明する。図 1 0 のステップ S 2 0 4 ～ S 2 0 6 はプローブセット選択処理の手順を示している。なお、必要があれば、プローブ候補を抽出して不適当なプローブを排除する。前述のプローブ評価ステップにおいて、ある範囲の長さ、または、ある範囲の融解温度のプローブのみを選んで得られたプローブ候補を評価したが、例えば、このプローブセット選択ステップにおいて、その判断を行うようにしてもよい。その場合、判別ツリー上の全てのプローブについて上述したステップ S 2 0 2 と S 2 0 3 のプローブ評価を行なうことになる。

## 【 0 0 5 0 】

ステップ S 2 0 4 において、プローブ候補となっている塩基配列の集合を複数のサブセットに分ける。これは冗長なプローブセットを選ぶことを防ぐためである。プローブの特異性を示す情報は、それぞれのターゲット塩基配列タイプに含まれるか含まれないかのビット列で表現できる。例えば、塩基配列タイプが  $t$  個存在するとすれば、 $t$  ビットの情報でプローブの特異性が表現できるわけである。塩基配列タイプが含む場合ビットを 1 に設定し、含まない場合ビットを 0 に設

定したとすると、例えば全てのタイプの塩基配列に含まれるプローブは全てが1であるtビットの値で表現されることとなる。このtビット列の異なるプローブを異なるグループとするとプローブの特異性を反映したサブセットが出来上がる。この処理によって同じサブセットに含まれる全てのプローブは、同タイプの特異性を持ち、情報的にはプローブとして同じ機能を持つこととなる。従って、各タイプのサブセットから上述したプローブ評価における評価値の高いプローブを選択してこれをマイクロアレイに用いるようにしてもよいが、本実施形態ではこのプローブセットを単位として更なる絞込を行なう。

## 【 0 0 5 1 】

上述のようにプローブ候補の塩基配列群を複数のサブセットに分けると、この複数のサブセットの中から必要でないサブセットを除くことが可能である。この処理を行うと、必要十分な独立プローブセットが出来上がる。以下に、具体的に例を示してこの処理を説明する。

## 【 0 0 5 2 】

ターゲットとなる塩基配列タイプがA、B、C、Dの4種類だったとする。いま、プローブ候補の塩基配列がその特異性に関して5つのサブグループに分けられるとすると、それぞれのサブグループの特異性が前述の4ビットの値で表せることとなる。ここで、例えば5つのサブグループが、1010,1100,0001,1110,0111というようなビット列でその特異性が表せたとする。すると実は、情報的には最初の2つのサブグループ(1010,1100)で4つのタイプを決定することは必要十分にできるのである。

## 【 0 0 5 3 】

この場合、最初の2つのサブグループの中から選んだプローブを $\alpha$ 、 $\beta$ とすると、もし試料がAタイプの塩基配列ならば、 $\alpha$   $\beta$ ともにハイブリダイゼーション反応する。同様に、試料がBタイプなら $\alpha$ にはハイブリダイゼーション反応は起こらず $\beta$ に対してのみ反応が起こる。試料がCタイプなら $\alpha$ にのみハイブリダイゼーション反応が起き、 $\beta$ には起きない。試料がDタイプなら $\alpha$ 、 $\beta$ ともにハイブリダイゼーション反応は起こらない。

## 【 0 0 5 4 】

もし実験が完全に制御できて、上記のハイブリダイゼーション反応の発生、非発生情報が非常に高い確率で得られたとすると、ターゲットとなるA、B、C、Dの4種類のタイプを判別する塩基プローブは、1番目と2番目のサブグループに含まれるプローブで必要十分となる。このような状況を情報的な観点でいうと、1番目と2番目のサブグループが独立で、その生成する情報空間に3番目、4番目、5番目のサブグループが従属しているという。よって、3番目、4番目、5番目のサブグループに属するプローブ候補は厳密に言えば必要なくなる。

## 【 0 0 5 5 】

なお、この必要十分な独立サブセットの組を選ぶ処理は非常に時間のかかる処理となる。また、実際は、最終的に選択するプローブセットは、ある程度冗長に選ぶ必要がある場合が多い。というのも、コンピュータの扱うデータと違って、生体物質の反応の結果として得られるデータは、実験の途中で多くのノイズが入る可能性が高く、必要最低限のプローブセットしか用意していないと、このノイズの多い場合に的確な実験結果を再現できないこととなるためである。よって、実際は上で述べたような必要十分な独立サブセットのみを選ぶ必要はない場合が多い。

## 【 0 0 5 6 】

ステップS 2 0 5では以上のようにしてプローブセットを選択し、最終的に、対象となるプローブ候補のサブセットを得る。そして、選択されたそれぞれのサブセットの中で、プローブ評価ステップで得られた評価値の高いプローブを最終的なプローブセットとして選択する。なお、本実施形態では、上記手順で得られる塩基配列はターゲットの塩基配列から抽出された部分塩基配列そのものの配列順序を有するので、ステップS 1 0 6で、選択された部分塩基配列を相補的な塩基配列に変換して出力することになる。例えば、ステップS 1 0 5で部分塩基配列「G A G C G」が選択されていれば、ステップS 1 0 6では、これに対応するプローブとして「C T C G C」が出力されることになる。もちろん、ステップS 1 0 2において判別ツリーを作成する時点で、ターゲットの塩基配列と相補的な塩基配列を用いてもよい。この場合は、ステップS 1 0 5で選択されるプローブの塩基配列がそのままステップS 1 0 6で出力されることになる。



## 【 0 0 5 7 】

以上のようにして選択されたプローブを用いることにより、DNAマイクロアレイシステムに最適なオリゴヌクレオチドプローブの設計が実現できる。そして、このことにより、より確かな遺伝子発現情報や、個体識別情報が得られるという効果がある。

なお、上記実施形態において説明した評価については、種々の変形が考えられる。例えば、評価関数によって求めた値が所定値を超える場合に、即座にそのノードに対応する部分塩基配列をプローブとして決定するように処理を簡略化することも可能である。或いは<数1>で算出したエントロピーの変化量が所定値を超える部分を有する部分塩基配列をプローブとして決定するように処理を簡略化することも可能である。

## 【 0 0 5 8 】

## &lt;第2の実施形態&gt;

第1の実施形態ではプローブ選択のために判別ツリーを生成し、これに基づいてプローブ評価を行なった。第2の実施形態では、判別ツリーの代わりにハッシュ表を用いる。以下、第2の実施形態の処理を説明する。なお、装置構成は第1の実施形態（図2）と同様であるので説明を省略する。

## 【 0 0 5 9 】

図11は第2の実施形態によるプローブ設計方法の処理手順を説明するフローチャートである。ステップS1001においてターゲットの塩基配列データを取得し、ステップS1002において、取得された塩基配列データに含まれるある長さの範囲の部分塩基配列を分類する部分配列ハッシュ表を作成する。そして、ステップS1003において、ステップS1002で作成された部分塩基配列ハッシュ表を外部記憶装置201に格納する。

## 【 0 0 6 0 】

ステップS1004は部分塩基配列ハッシュ表に存在するプローブ候補について、プローブとしてふさわしいかどうかを評価するプローブ評価処理を実行する。ステップS1005では、更にプローブセットの選択を行ない、上記プローブ評価の結果を元に最終結果の最適プローブを選択する。ステップS1006は、

最終結果として得られた最適プローブである。上記処理において、入力ターゲットとなる塩基配列データであり、出力はマイクロアレイに搭載するのに適したプローブである。

#### 【0061】

〔部分塩基配列ハッシュ表の生成について（S1002、S1003）〕

第2の実施形態では、第1の実施形態で説明した判別ツリーを作成する替わりに、部分塩基配列ハッシュ表を作成する。これは判別ツリーのある深さのノードを作成するのに対応する。ハッシュ表とはソフトウェア技術として非常によく使われる方法で、大量のバリエーションのありうる情報列をコンパクトに保存する方法である。

#### 【0062】

例えば、20ベースの長さの部分塩基配列がターゲットとなる塩基配列の中にどのように含まれるかを見る場合を考えてみる。20ベースの塩基配列の取りうるバリエーションは、理論的には $4^{20}$ （＝40ビット）となる。これだけの大きな空間を用意することは事実上不可能である。一方、試料中に存在しうる塩基配列のバリエーションは、この40ビットの空間と比べると、極めて小さい。よって、20ベースの部分塩基配列を40ビットの値にコーディングしておいて、それを例えば十分大きな素数でわることによって、事実上重複のないコンパクトな空間に収めることができる。この方法をハッシュ表を用いたデータ格納整理方法と呼んでいる。なお、仮に実際は異なる塩基配列にもかかわらず、重複が生じた場合は、2重にハッシュ表を用意するとか、ハッシュ表の次の番地に格納するというような回避方法がある。

#### 【0063】

ハッシュ表を用いて、ターゲットの塩基配列の中の特定の長さの全部分配列が格納整理できると、それが含まれる塩基配列の種類と位置も自動的に特定できる。これは、図7や図8の左のカラムがハッシュ表のそれぞれのエントリーに登録されている状態と同じである。このことによって、実施形態1の判別ツリーのある深さのノード情報が全て得られたことになる。

#### 【0064】

この操作を複数の長さの部分塩基配列に対して実施することによって、第 1 の実施形態である深さからある深さまでの全てのノード展開ができたことと同じになる。第 1 の実施形態で述べたとおり、ある範囲の長さのプローブやある範囲の融解温度のプローブを設計する場合は、これでプローブ候補が全て得られたこととなる。

## 【 0 0 6 5 】

## 〔プローブ候補の評価について (S 1 0 0 4) 〕

次に、この全てのプローブ候補から、プローブとして適切か否かを判定するためのプローブ評価について説明する。第 2 の実施形態が第 1 の実施形態と決定的に異なるところは、このプローブ評価処理である。第 2 の実施形態は第 1 の実施形態より若干性能が劣る。つまり、第 1 の実施形態の方がより高性能なプローブを設計できる場合が多い。しかしながら、第 2 の実施形態は、処理の簡易さ、迅速さにおいて第 1 の実施形態を上回るものであり、その利用価値は高いものである。

## 【 0 0 6 6 】

第 1 の実施形態でも述べたとおり、プローブの塩基配列の特異性は中央部分に存在するほうがプローブとしては質が高い。よって、その中央部の特異性を調べたいのであるが、第 1 の実施形態と違って、塩基配列中の部分的な情報が全くないので非常に困難である。そこで、第 2 の実施形態では、ターゲットの塩基配列中に存在する特異点（特異塩基位置という）を予め求める。この様子を示したのが図 1 2 である。図 1 2 の表の一番下に表示したのが特異塩基位置である。\* のマークのある場所が、ターゲットの塩基配列の全てが共通とはなっていない箇所、すなわちあるタイプ間に特異性がある場所（特異塩基位置）である。そして、この位置を中央に含むプローブ候補が良質なプローブとなる。この特異塩基位置のプローブ候補中での場所を評価する関数を使って、プローブ候補の質を評価する。

## 【 0 0 6 7 】

具体的には、例えば、

## 【数 4】

評価スコア =  $\sum \{1 / (\text{中心位置からの特異塩基位置の距離} + 1)\}^n$

というような評価関数を利用して判定することができる。

【0068】

また、上述の例では、特異塩基位置の判定において、特異性があるかないかの2値で評価していたが、実施形態1に示したようなエントロピーのような尺度を使用して、連続値で評価しても問題ない。例えば、図12の\*のマークのあるところは、特異塩基配列と呼ばれるが、場所によってはほとんどがGで、一種類のタイプだけAである場所（エントロピーの減少量が低い）と、半分のタイプがGで、残りの半分のタイプがAである場所（エントロピーの減少量が高い）が混じっている。上述したように、エントロピーの減少量が多い場所が中央にあったほうが判別に有利なので、＜数4＞に対してそのような重み付けを行なうこと、例えば、特異塩基配列の位置におけるエントロピーの減少量を＜数4＞に積算するというような重み付けをすることで、より正確なプローブ候補の質評価が可能となる。

【0069】

〔プローブセット選択処理（S1005）〕

ステップS1005で実行されるプローブセット選択処理は、第1の実施形態（ステップS106）と同様の処理を行うものであり、説明を省略する。

【0070】

以上のように第2の実施形態によれば、ハッシュ表を用いてプローブ候補を選択するので、処理の簡素化、高速化が達成される。

【0071】

以上説明したように、第1或いは第2の実施形態のプローブ設計処理に従って得られた塩基プローブを基板上に配置することで、ターゲットのタイピングに好適な塩基プローブを含むDNAマイクロアレイが得られることになる。なお、塩基プローブが決定されれば、周知のDNAマイクロアレイの製造方法を用いて、所定のターゲットに対して好適なDNAマイクロアレイを製造することができる。また、このようなDNAマイクロアレイを用いる遺伝子検査装置を提供することで、ターゲットのタイピングを良好に実現することができる。なお、DNAマ

マイクロアレイの製造方法と構造、およびこれを用いた遺伝子検査装置そのものは周知であるので、その詳細な説明を省略する。

#### 【0072】

##### <第3実施形態>

上述の各実施形態においては、PCRなどの手法を用いて、調べたいDNA、あるいは、RNA（cDNA）の塩基配列を増幅して検体溶液を作成するのが、最も一般的であった。この場合、検体中に存在する塩基配列は、調べたい塩基配列をその部分配列として包含している可能性が非常に高かった。即ち、第1及び第2実施形態は、PCR等の実験手法を用いた検体調製をした場合に、特に好適である。

#### 【0073】

ところが、例えば細胞中に存在するRNAをまとめて抽出し、その中の一部のRNAの存在量を測定するといった目的の場合、作成したDNAマイクロアレイと反応する可能性のある塩基配列は、細胞中に存在する全てのRNAとなる。そして、これらのRNA、あるいはそれから作成されたcDNAの塩基配列の中に調べたい塩基配列が存在する確率は極めて低い。

#### 【0074】

これを図を使って説明すると、図13のようになる。第1及び第2実施形態では、「検体中に存在する可能性のある塩基配列、あるいはその部分塩基配列」と、「調べたい塩基配列、あるいはその部分塩基配列」は、ほとんど同じ集合であった。これを第1及び第2実施形態においては、ターゲット塩基配列と呼んでいた。例えば、MHCなどの人体特異性を示す遺伝子DNAの解析などは、この場合に当てはまる。

#### 【0075】

これに対して、上述のようなRNAの発現量を測定する実験の場合、検体中に存在する塩基配列は、調べたい塩基配列に対して、圧倒的に種類が多いのが一般的である。よって、この場合は、検体中に存在する可能性のある全ての塩基配列が特異性の対象となるターゲット塩基配列となる。そして、調べたい塩基配列は、そのターゲット塩基配列の一部となる。

## 【 0 0 7 6 】

図 1 3 で説明すると、第 1 実施形態の判別ツリー、または、第 2 実施形態のハッシュテーブルを作成するに際して、使用される塩基配列は、図 1 3 の「検体中に存在する可能性のある塩基配列の全部分塩基配列」(1 3 0 1)となる。そして、調べたい塩基配列の部分配列の特異性は、全体であるターゲット塩基配列との特異性に基づいて導かれ、プローブ候補が決定されることとなる。

## 【 0 0 7 7 】

以下、図 1 4、図 1 5 を用いて、第 3 実施形態によるプローブ設計方法を具体的に説明する。

## 【 0 0 7 8 】

まず、第 1 実施形態と同様の判別ツリーを用いたプローブ設計手順について図 1 4 を用いて説明する。ステップ S 1 4 0 1 において、検体中に存在する可能性のある全塩基配列であるターゲット塩基配列 1 4 1 1 を取得する。この中には、調べたい塩基配列データ 1 4 1 0 が含まれる。ターゲット塩基配列 1 4 1 1 は、図 1 3 に示されるように、調べたい塩基配列 1 3 0 2 を含む塩基配列集合 1 3 0 1 である。

## 【 0 0 7 9 】

ステップ S 1 4 0 2 では、ステップ S 1 4 0 1 で得られたターゲット塩基配列データ 1 4 1 1 より得られる複数の部分塩基配列を分類する判別ツリー 1 4 1 2 を生成する。判別ツリーの生成については第 1 実施形態(ステップ S 1 0 2、S 1 0 3)で説明したとおりである。

## 【 0 0 8 0 】

次に、ステップ S 1 4 0 3 において、判別ツリー 1 4 1 2 の中に存在するプローブ候補についてプローブとしてどの程度ふさわしいかを評価する。評価の方法は第 1 の実施形態(ステップ S 1 0 4)と同様である。但し、プローブとして選ばれる部分塩基配列は、調べたい塩基配列データ 1 4 1 0 の部分配列の中から選ばれ、その特異性は、ターゲット塩基配列データ 1 4 1 1 をもとに求められることとなる。

## 【 0 0 8 1 】

続いて、ステップ S 1 4 0 4 において、ステップ S 1 4 0 3 におけるプローブの評価結果に基づいて最終結果の最適プローブセットを選択する。この処理は第 1 の実施形態のステップ S 1 0 5 と同様である。そして、ステップ S 1 4 0 5 において、ステップ S 1 4 0 4 で選択されたプローブが最適プローブとして出力される。本実施形態の入力は、調べたい塩基配列データであり、出力は調べたい塩基配列に含まれる最適プローブセット、またはその相補鎖となる。

#### 【 0 0 8 2 】

次に、第 2 実施形態と同様のハッシュテーブルを用いたプローブ設計手順について図 1 5 を参照して説明する。

#### 【 0 0 8 3 】

ステップ S 1 5 0 1 において、検体中に存在する可能性のある全塩基配列であるターゲット塩基配列 1 5 1 1 を取得する。この中には、調べたい塩基配列データ 1 5 1 0 が含まれる。ターゲット塩基配列 1 5 1 1 は、図 1 3 に示されるように、調べたい塩基配列 1 3 0 2 を含む塩基配列集合 1 3 0 1 である。

#### 【 0 0 8 4 】

ステップ S 1 5 0 2 では、ステップ S 1 5 0 1 で得られたターゲット塩基配列データ 1 5 1 1 より得られる複数の部分塩基配列を分類する部分塩基配列ハッシュ表 1 5 1 2 を生成する。部分塩基配列ハッシュ表 1 5 1 2 の生成については第 2 実施形態（ステップ S 1 0 0 2、S 1 0 0 3）で説明したとおりである。

#### 【 0 0 8 5 】

次に、ステップ S 1 5 0 3 において、部分塩基配列ハッシュ表 1 5 1 2 の中に存在するプローブ候補についてプローブとしてどの程度ふさわしいかを評価する。評価の方法は第 2 の実施形態（ステップ S 1 0 0 4）と同様である。但し、プローブとして選ばれる部分塩基配列は、調べたい塩基配列データ 1 5 1 0 の部分配列の中から選ばれ、その特異性は、ターゲット塩基配列データ 1 5 1 1 をもとに求められることとなる。

#### 【 0 0 8 6 】

続いて、ステップ S 1 5 0 4 において、ステップ S 1 5 0 3 におけるプローブの評価結果に基づいて最終結果の最適プローブセットを選択する。この処理は第

2の実施形態のステップS1005と同様である。そして、ステップS1505において、ステップS1504で選択されたプローブが最適プローブとして出力される。本実施形態の入力は、調べたい塩基配列データであり、出力は調べたい塩基配列に含まれる最適プローブセット、またはその相補鎖となる。

【0087】

以上のように、本実施形態によれば、例えば細胞中に存在するRNAをまとめて抽出し、その中の一部のRNAの存在量を測定するといった用途に対して、適切なマイクロアレイを設計することが可能となる。

【0088】

また、本発明の目的は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはCPUやMPU）が記憶媒体に格納されたプログラムコードを読み出し実行することによっても、達成されることは言うまでもない。

【0089】

この場合、記憶媒体から読出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。

【0090】

プログラムコードを供給するための記憶媒体としては、例えば、フロッピディスク、ハードディスク、光ディスク、光磁気ディスク、CD-ROM、CD-R、磁気テープ、不揮発性のメモ리카ード、ROMなどを用いることができる。

【0091】

また、コンピュータが読出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているOS（オペレーティングシステム）などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【0092】



さらに、記憶媒体から読出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるCPUなどが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【 0 0 9 3 】

【発明の効果】

以上説明したように、本発明によれば、解析対象となるターゲットの塩基配列に応じて、解析に適切なプローブを自動的に選択することが可能となり、プローブ設計の効果的な支援が実現される。

【 0 0 9 4 】

【配列表】

<110> CANON INC.

<120> Probe Design Method And Information Processing Apparatus

<140> 2001-055465

<141> 2001-02-28

<160> 8

<210> 1

<211> 50

<212> DNA

<213> Homo Sapiens

<400> 1

gagcgggtgc ggttgctgga aagatgcatc tataaccaag aggagtccgt 50

<210> 2

<211> 50

<212> DNA

<213> Homo Sapiens

<400> 2

gagcgggtgc ggttcctgga cagatacttc taccaccaag aggagtacgt 50

<210> 3

<211> 50

<212> DNA

<213> Homo Sapiens

<400> 3

gagcgggtgc agttcctgga aagactcttc tataaccagg aggagttcgt 50

<210> 4

<211> 50

<212> DNA

<213> Homo Sapiens

<400> 4

gagcgggtgc ggttcctgga cagatacttc tataaccagg aggagtccgt 50

<210> 5

<211> 49

<212> DNA

<213> Homo Sapiens

<400> 5

agcgggtgcg gttgctggaa agatgcattc ataaccaaga ggagtccgt 49

<210> 6

<211> 48

<212> DNA

<213> Homo Sapiens

<400> 6

gcgggtgcgg ttgctggaaa gatgcatcta taaccaagag gagtccgt 48

<210> 7

<211> 47

<212> DNA

<213> Homo Sapiens

<400> 7

cgggtgcggt tgctggaaaag atgcatctat aaccaagagg agtccgt 47

<210> 8

<211> 46

<212> DNA

<213> Homo Sapiens

<400> 8

gggtgcggtt gctggaaaga tgcatctata accaagagga gtccgt 46

【図面の簡単な説明】

【図 1】

第 1 の実施形態によるプローブ設計方法の処理手順を説明するフローチャートである。

【図 2】

本実施形態によるプローブ設計方法が適用される情報処理装置の構成を示すブロック図である。

【図 3】

ステップ S 1 0 1 で取得する塩基配列データの例を示す図である。

【図 4】

DNA マイクロアレイ上のハイブリダイゼーションの様子を示した図である。

【図 5】

ターゲットとなる塩基配列から得られる部分塩基配列（プローブ）を示す図で

ある。

【図 6 A】

第 1 の実施形態における判別ツリーを説明する図である。

【図 6 B】

判別ツリーにおける各ノードの内容例を示す図である。

【図 7】

判別ツリー上における、あるノードに登録された部分塩基配列の様子を示す図である。

【図 8】

判別ツリー上における、あるノードに登録された部分塩基配列の様子を示す図である。

【図 9】

プローブとして典型的に最適なエントロピーの推移をグラフに示した図である。

【図 1 0】

第 1 の実施形態によるプローブの評価処理（ステップ S 1 0 4）とプローブセットの選択処理（ステップ S 1 0 5）の詳細を示すフローチャートである。

【図 1 1】

第 2 の実施形態によるプローブ設計方法の処理手順を説明するフローチャートである。

【図 1 2】

第 2 の実施形態による、ターゲット中の特異塩基位置を説明する図である。

【図 1 3】

第 3 の実施形態によるプローブ設計方法が適用される塩基配列データの状態を説明する図である。

【図 1 4】

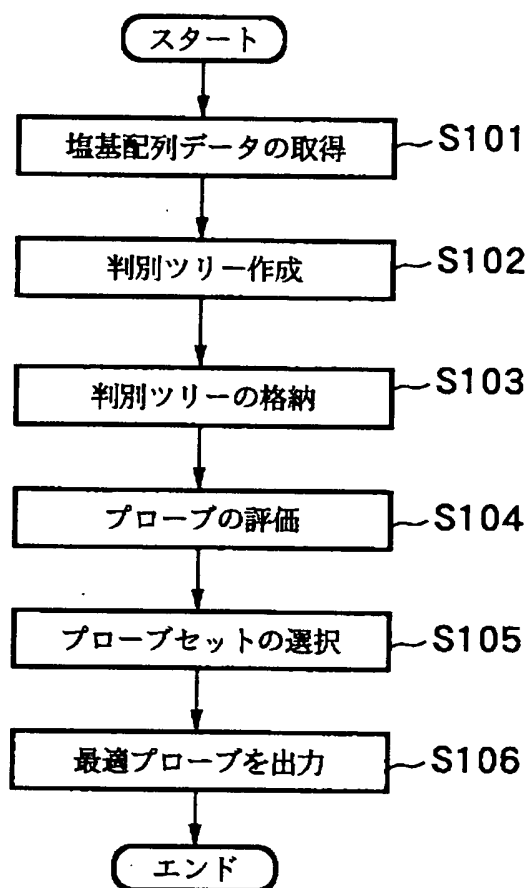
第 3 の実施形態による、判別ツリーを用いたプローブ設計方法を説明するフローチャートである。

【図 1 5】

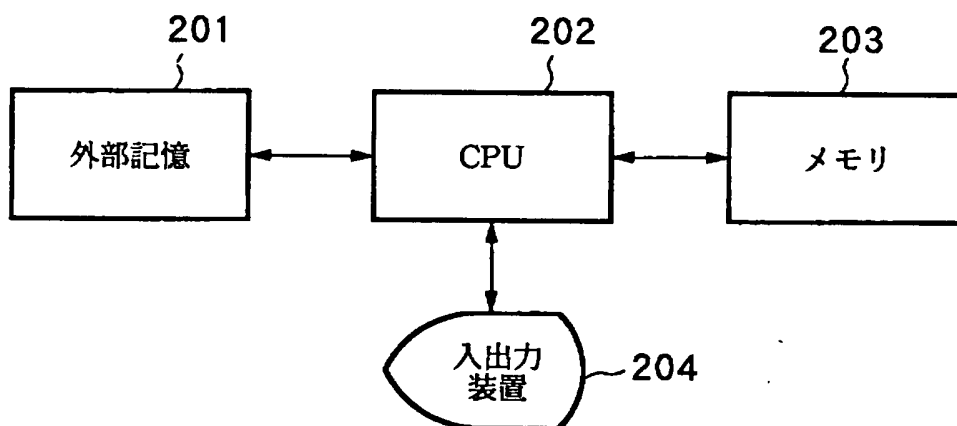
第 3 の実施形態による、ハッシュ表を用いたプローブ設計方法を説明するフローチャートである。

【書類名】 図面

【図 1】



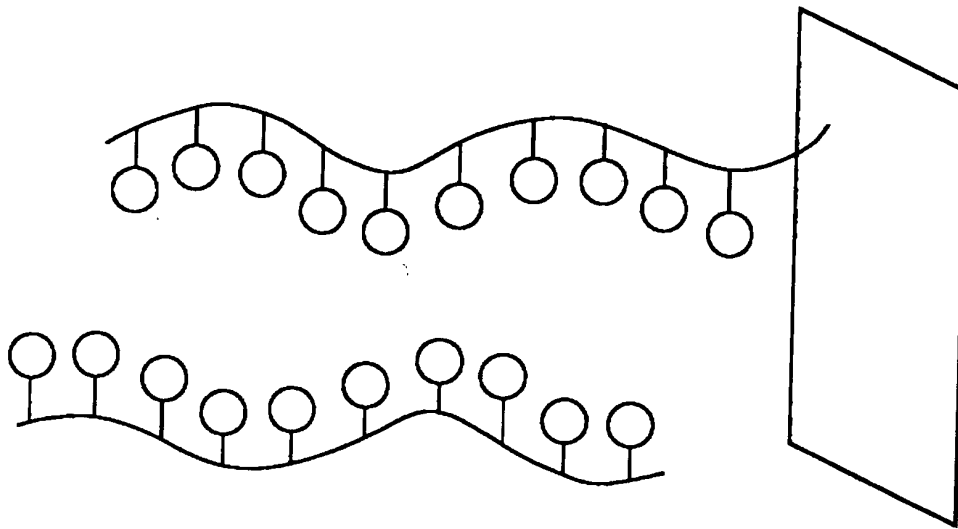
【図 2】



【図 3】

DRB1*0101	GAGCGGCTGC GGTGCTGGA AAGATGCATC TATAACCAAG AGGACTCCGT ... ..
DRB1*04011	GAGCGGCTGC GGTCTCTGGA CAGATACTTC TATCACCAGG AGGACTACGT ... ..
DRB1*07011	GAGCGGCTGC AGTCTCTGGA AAGACTCTTC TATAACCAGG AGGACTTCGT ... ..
DRB1*15011	GAGCGGCTGC GGTCTCTGGA CAGATACTTC TATAACCAGG AGGACTCCGT ... ..
...	... ..

【図4】

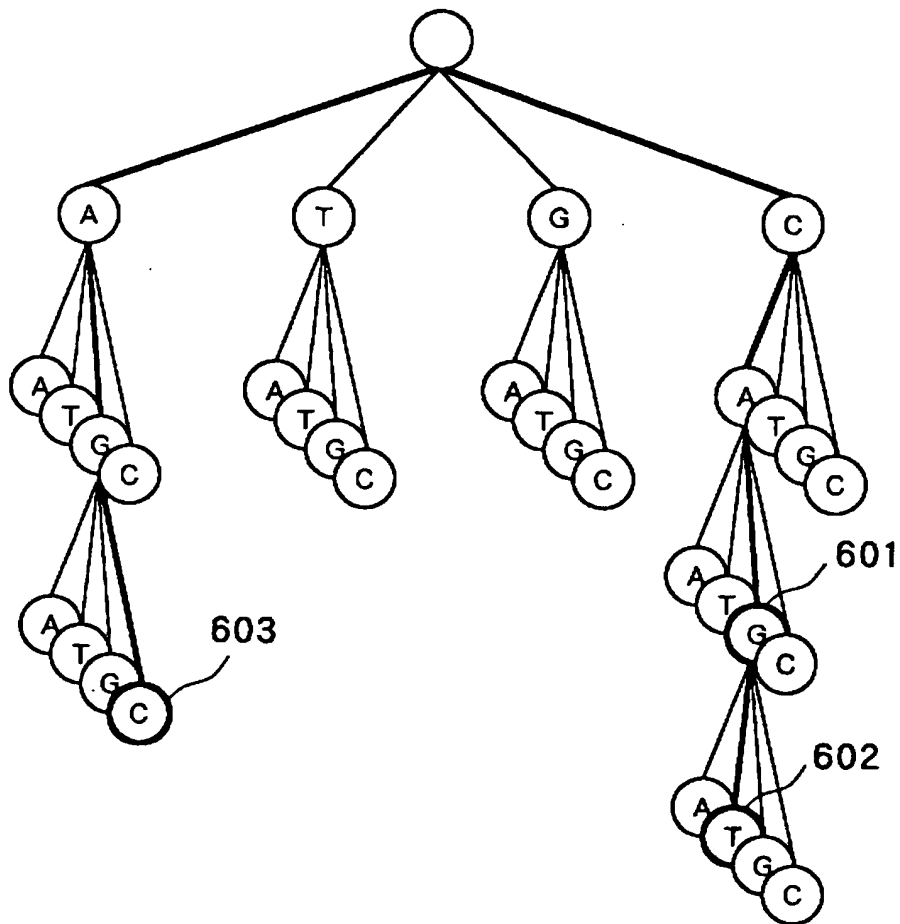




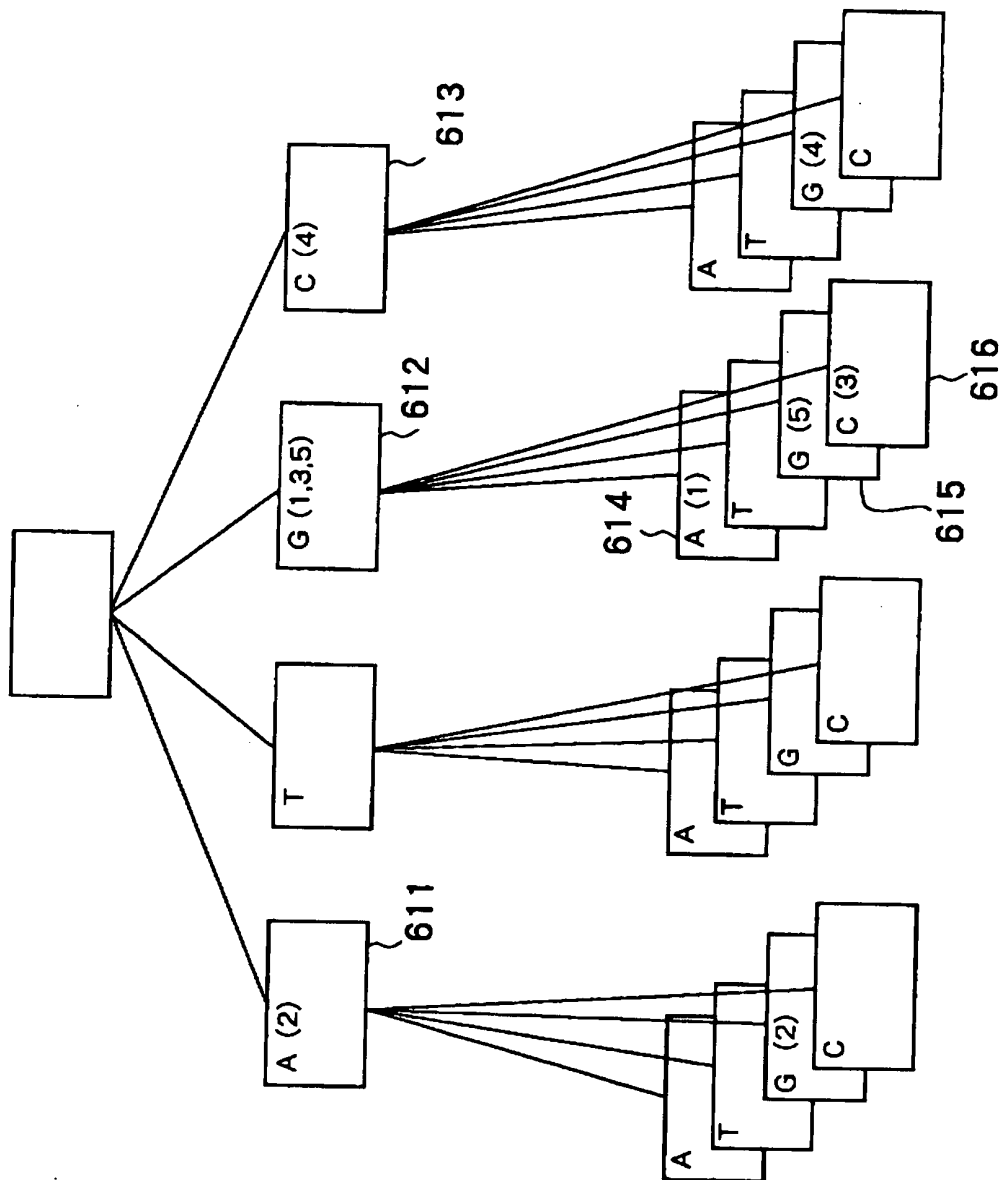
【図 5】

1	GAGCGGTGC GGTGCTGGA AAGATGCATC TATAACCAAG AGGAGTCCGT ... ..
2	ACCGGTGC GGTGCTGGA AAGATGCATC TATAACCAAG AGGAGTCCGT ... ..
3	GCGGTGC GGTGCTGGA AAGATGCATC TATAACCAAG AGGAGTCCGT ... ..
4	CGGTGC GGTGCTGGA AAGATGCATC TATAACCAAG AGGAGTCCGT ... ..
5	GGTGC GGTGCTGGA AAGATGCATC TATAACCAAG AGGAGTCCGT ... ..
::	::

【図 6 A】



【図6B】



【図 7】

ノードに登録される部分配列 CAG

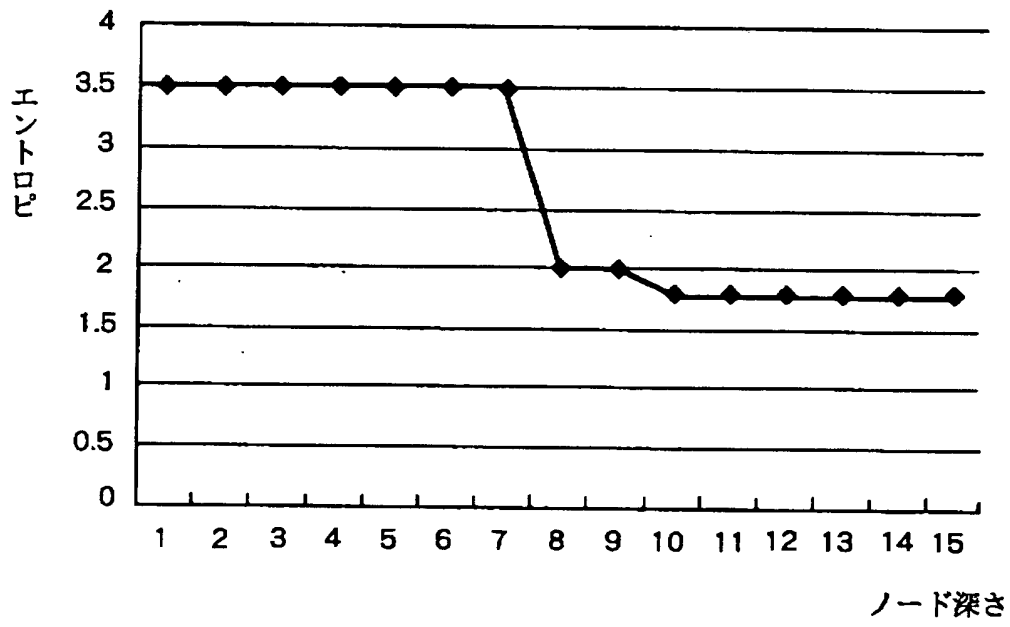
DRB1*0101	GAGCGGCTGC GGTGCTGGA AAGATGCATC TATAACCAAG AGGAGTCCGT	NULL
DRB1*04011	GAGCGGCTGC GGTTCCTGGA <u>CAG</u> ATACTTC TATCACCAG AGGACTACGT	21
DRB1*07011	GAGCGGCTGC <u>AG</u> TTTCCTGGA AAGACTCTTC TATAA <u>CAGG</u> AGGACTTCGT	10, 37
DRB1*15011	GAGCGGCTGC GGTTCCTGGA <u>CAG</u> ATACTTC TATAA <u>CAGG</u> AGGAGTCCGT	21, 37

【図 8】

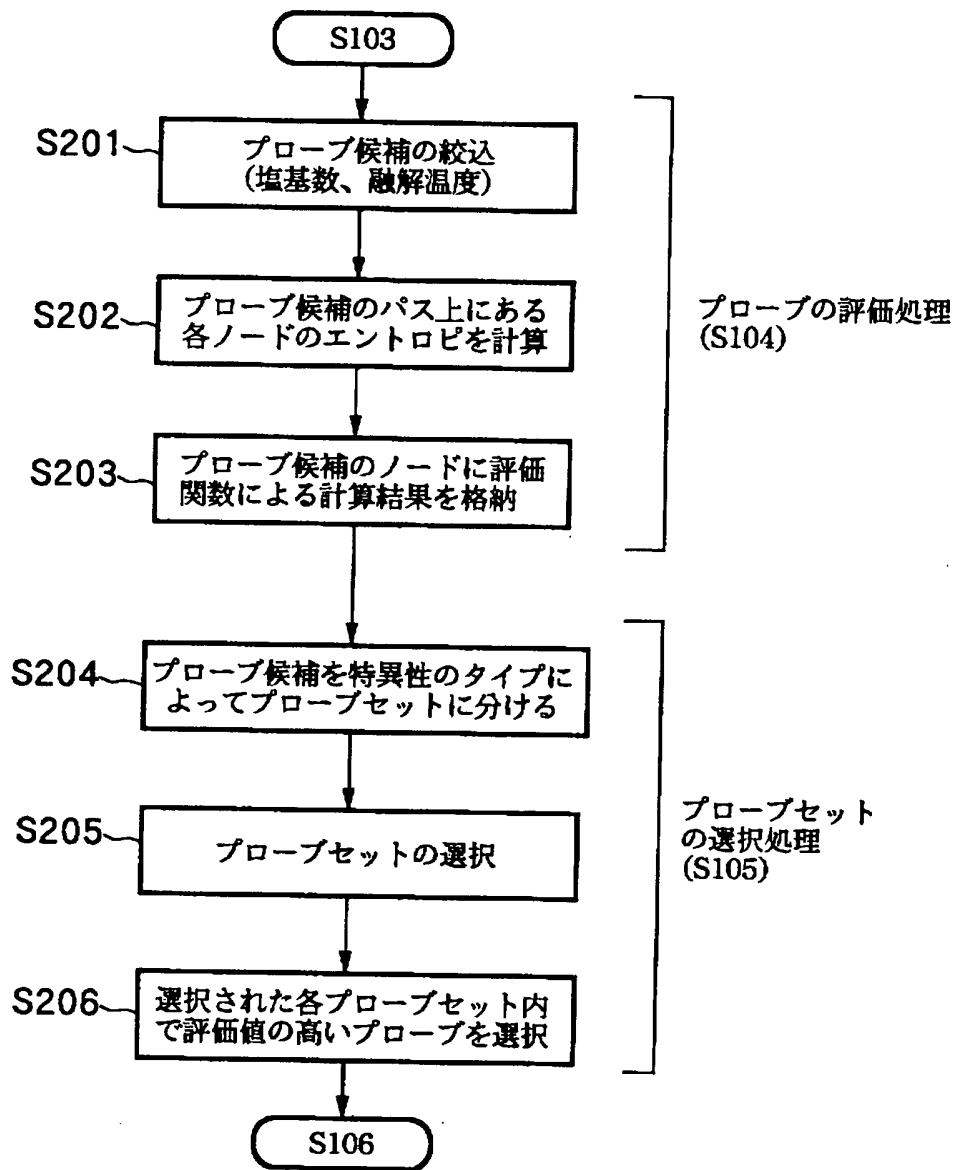
ノードに登録される部分配列 CAGT

DRB1*0101	GAGCGGTGC GGTTCCTGGA AAGATGCATC TATAACCAAG AGGAGTCCGT	NULL
DRB1*0401	GAGCGGTGC GGTTCCTGGA CAGATACTTC TATCACCAGG AGGAGTACGT	NULL
DRB1*0701	GAGCGGTGC <b>AGT</b> TCCTGGA AAGACTCTTC TATAACCAGG AGGAGTTCGT	10
DRB1*1501	GAGCGGTGC GGTTCCTGGA CAGATACTTC TATAACCAGG AGGAGTCCGT	NULL

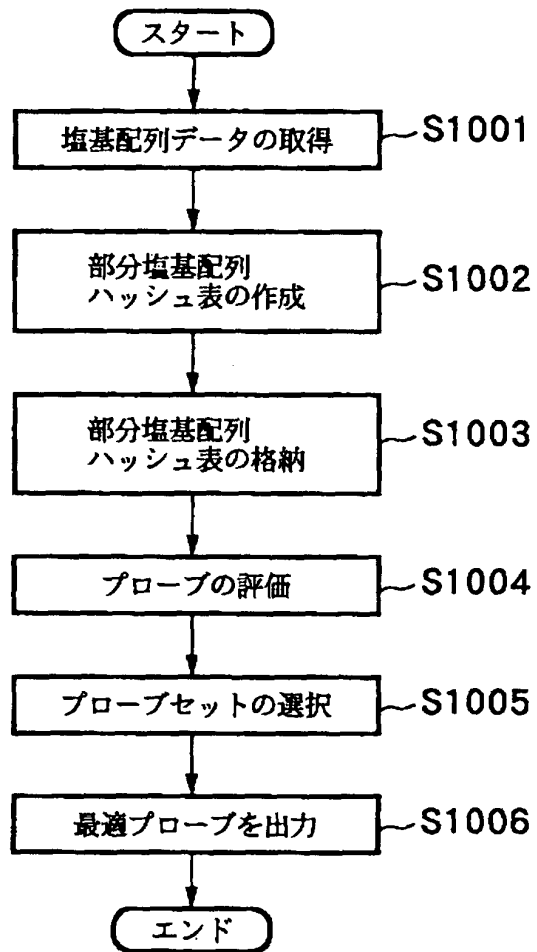
【図9】



【図 1 0】



【図 11】

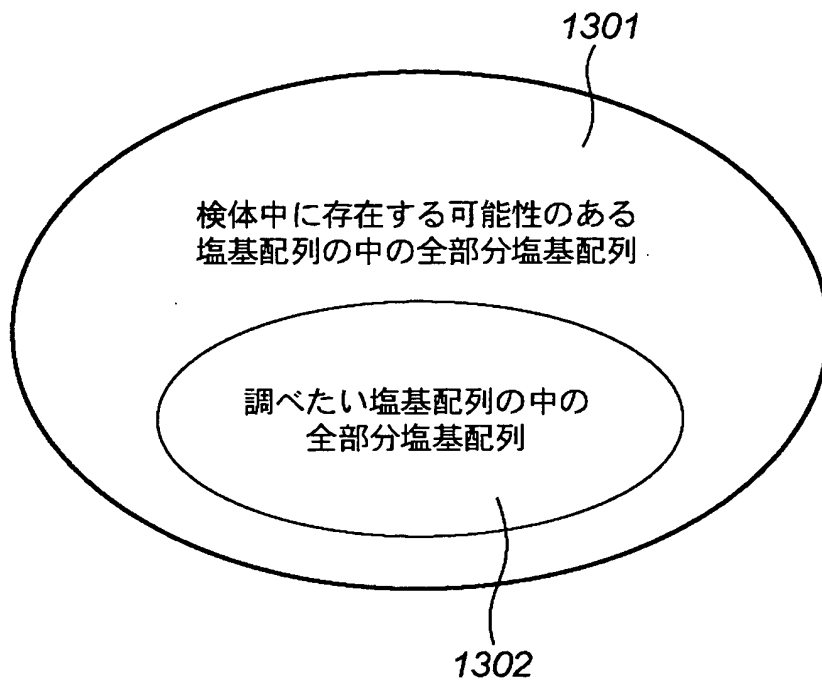




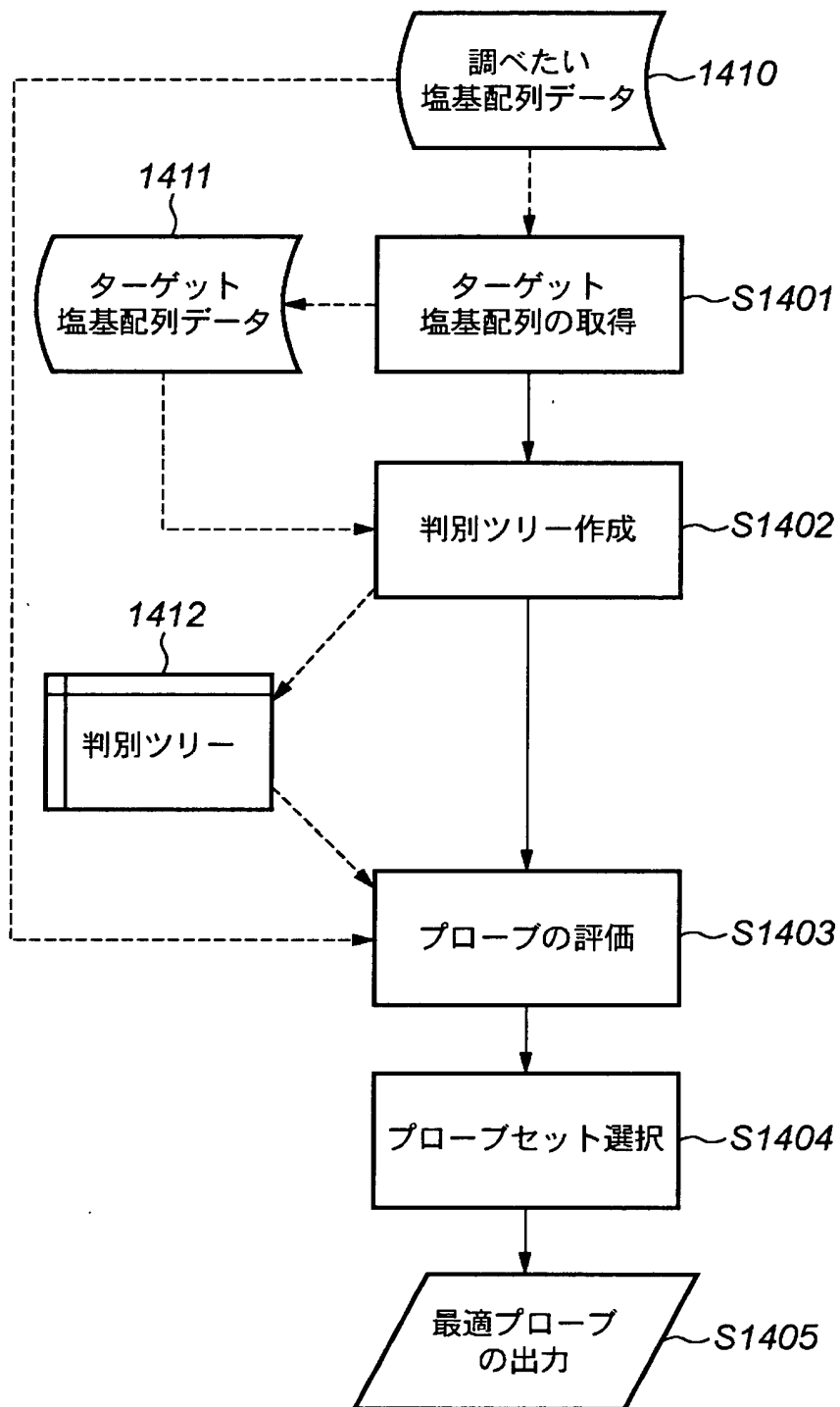
【図12】

DRB1*0101	GAGCGGGTGC GCTTGCTGGA AAGATGCAIC TATAACCAAG AGGAGTCCGT
DRB1*04011	GAGCGGGTGC GCTTCCTGGA CAGATACTTC TATCACCAAG AGGAGTACGT
DRB1*07011	GAGCGGGTGC AGTTCCTGGA AAGACTCTTC TATAACCAGG AGGAGTTCGT
DRB1*15011	GAGCGGGTGC GCTTCCTGGA CAGATACTTC TATAACCAGG AGGAGTCCGT
特異塩基位置	-----*-----*-----*-----*-----*

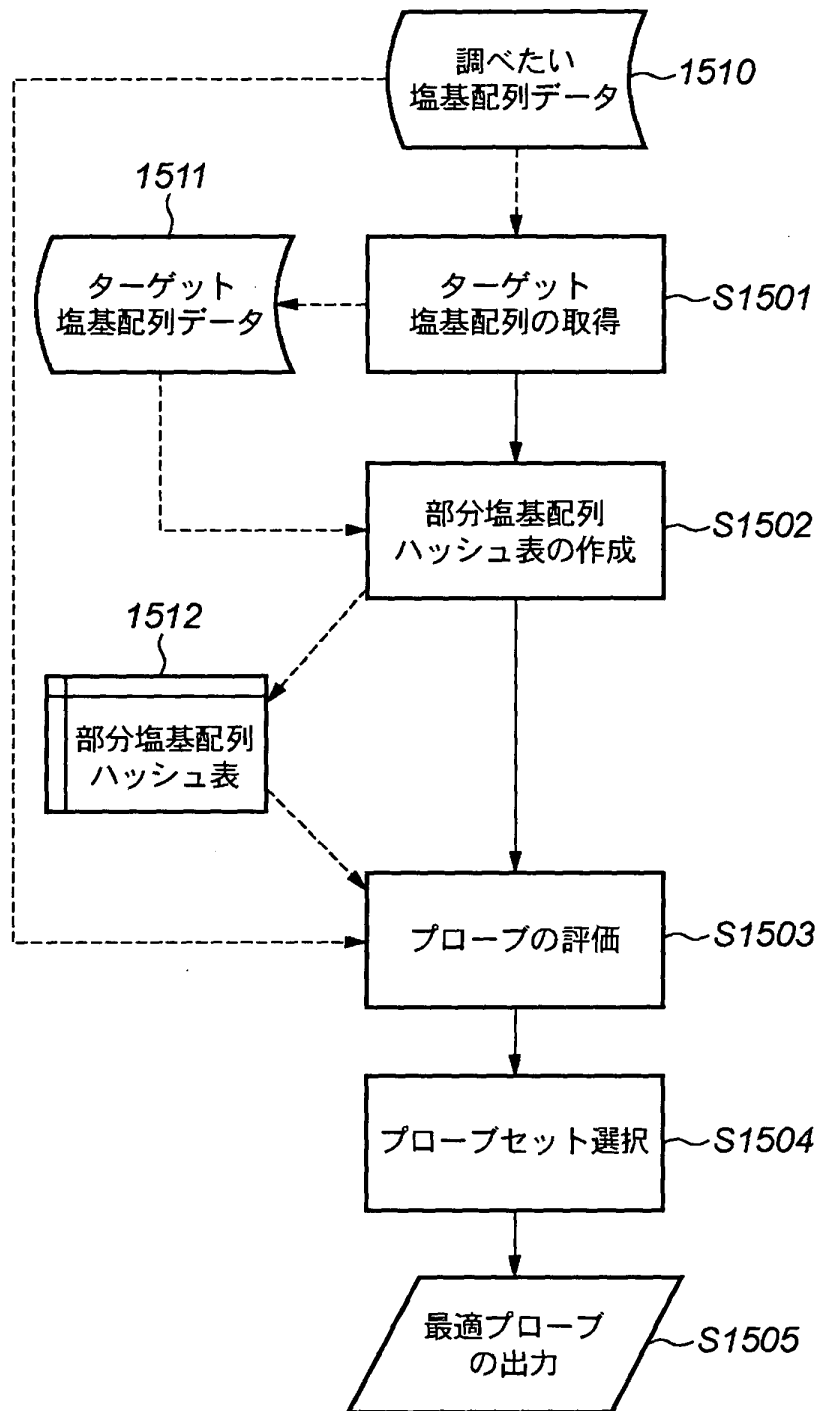
【図 1 3】



【図14】



【図15】



【書類名】 要約書

【要約】

【課題】 解析対象となるターゲットの塩基配列に応じて、解析に適切なプローブを自動的に選択することを可能とし、プローブ設計を効果的に支援する。

【解決手段】 未知の核酸断片とハイブリダイゼーション反応させて遺伝子解析を行うためのプローブに用いられる塩基配列を自動設計するにおいて、ターゲットの塩基配列データより得られる複数の部分塩基配列を各ノード上に配置するツリーを生成する（S101～S103）。次に、所望のノードに関して、該所望のノードが表す部分塩基配列のプローブとしての適性を評価する（S104）。また、複数の部分塩基配列について特異性に基づくグループ化を行ない、適切なグループの組み合わせを選択する（S105）。選択されたグループに属する部分塩基配列と、それらの上記評価結果に基づいてプローブとして使用すべき部分塩基配列を決定する（S106）。

【選択図】 図1

## 認定・付加情報

特許出願の番号	特願 2002-023953
受付番号	50200131709
書類名	特許願
担当官	小野寺 光子 1721
作成日	平成14年 2月 6日

### <認定情報・付加情報>

#### 【特許出願人】

##### 【識別番号】

000001007

##### 【住所又は居所】

東京都大田区下丸子3丁目30番2号

##### 【氏名又は名称】

キヤノン株式会社

#### 【代理人】

申請人

##### 【識別番号】

100076428

##### 【住所又は居所】

東京都千代田区紀尾井町3番6号 秀和紀尾井町  
パークビル7F 大塚国際特許事務所

##### 【氏名又は名称】

大塚 康德

#### 【選任した代理人】

##### 【識別番号】

100112508

##### 【住所又は居所】

東京都千代田区紀尾井町3番6号 秀和紀尾井町  
パークビル7F 大塚国際特許事務所

##### 【氏名又は名称】

高柳 司郎

#### 【選任した代理人】

##### 【識別番号】

100115071

##### 【住所又は居所】

東京都千代田区紀尾井町3番6号 秀和紀尾井町  
パークビル7F 大塚国際特許事務所

##### 【氏名又は名称】

大塚 康弘

#### 【選任した代理人】

##### 【識別番号】

100116894

##### 【住所又は居所】

東京都千代田区紀尾井町3番6号 秀和紀尾井町  
パークビル7F 大塚国際特許事務所

##### 【氏名又は名称】

木村 秀二

次頁無

出 願 人 履 歴 情 報

識別番号 [000001007]

1. 変更年月日 1990年 8月30日

[変更理由] 新規登録

住 所 東京都大田区下丸子3丁目30番2号

氏 名 キヤノン株式会社